

AUTOMATIC MORPHOLOGICAL PROCESSING OF BULGARIAN PROPER NOUNS

HRISTO KRUSHKOV*

Abstract – Résumé

This paper presents (i) a classification of Bulgarian proper nouns, (ii) a methodology for automatic morphological analysis and generation of proper nouns, (iii) some approaches to automatically build a dictionary of proper nouns. Bulgarian proper nouns are divided into classes. Every class comprises rules for generation of the paradigm. The pattern is a lexical representation, which matches all forms of the paradigm. The morphological analysis is based on the pattern matching process between the proper noun and the pattern. The pattern and the class incorporate information about the whole paradigm of a particular proper noun. An electronic dictionary of proper nouns has been created. It consists of pairs <pattern, class number>.

Cet article présente (i) une classification des noms propres bulgares, (ii) une méthodologie d'analyse et génération automatique des formes des noms propres, (iii) quelques approches à la création automatique d'un dictionnaire des noms propres. Les noms propres bulgares sont divisés en classes. Chaque classe comporte des règles pour la génération du paradigme. Le patron est une représentation lexicologique qui apparie toutes les formes du paradigme. L'analyse morphologique est basée sur l'appariement du nom propre et du patron. Le patron et la classe incorporent des informations sur le paradigme entier d'un nom propre particulier. Un dictionnaire électronique des noms propres est créé. Il contient paires <patron, numero de la classe>.

Keywords – Mots-clefs:

Bulgarian proper nouns, morphological analysis, morphological generation, electronic dictionaries.

noms propres bulgares, analyse morphologique, génération morphologique, dictionnaires électroniques.

* Department of Computer Science, University of Plovdiv
24 Tzar Assen Str, Plovdiv, Bulgaria. E-mail: hdk@pu.acad.bg

INTRODUCTION

There is a rapidly growing need for development of computer processing tools for natural language (NL), as well as for corresponding technologies on national and world markets. This need is strongly motivated by the spread of the Internet worldwide and the increase in the degree of computerisation of public and private structures in management and education.

Bearing in mind the complexity of creating such missing tools for the Bulgarian language and taking into account the world experience, the department of Computer Science at the University of Plovdiv, in cooperation with the Department of Bulgarian language at the same university and the unit for Linguistic Modelling at the Institute of Bulgarian Language at the Bulgarian Academy of Science started researches in the field of Computational Linguistics. We started out with the following three tasks:

- To develop a morphological processor.
- To compile a General Computer Dictionary as well as other linguistic resources of the Bulgarian language.
- To develop proofing tools for Bulgarian.

In order to solve these tasks, three projects in the field of Computational Linguistics were launched, funded by the National Fund **Scientific Researches**. Our global idea was to increase the speed of research through automatization of the processes of extracting, building and updating of linguistic resources and computer tools. In the last years we proposed effective techniques and computer tools that realised this idea. Methods for research of text corpora and automatic extraction of unknown grammatical features and structures have been proposed and applied (Totkov G. 1990, Totkov G. 1991, Tanev Ch. & Krushkov Hr. 2000). The developed techniques and tools are applicable to other natural languages too.

A morphological processor is a computational tool, which performs automatic morphological analysis and synthesis of wordforms using an electronic dictionary of base forms. The development of a morphological processor requires a detailed study of morphological phenomena. Inflectional morphology is the study of the way words change when used in different grammatical contexts, derivational morphology is concerned with the principles behind the creation of completely new words, and is less concerned with the grammatical role they play. Derivational rules tend to be much less productive, or regular, than inflectional rules, and they also are much more likely to involve changes in the category of a word than inflectional rules.

A profound introduction to the field of computational morphology is made in (Sproat R. 1992), where an integrated set of techniques for the description of morphological phenomena is presented. One popular way of specifying morphological properties of a language is with two-level rules (Koskenniemi K. 1983). They directly relate the visible surface form of a word to an abstract, lexical form. Two-level rules are usually implemented using finite-state automata or machines, in which the lexical and surface forms are compared.

One of the aims of the present study has been to analytically examine Bulgarian proper nouns in order to develop their morphological model. In this model the surface form (like in the two level model) relates to an abstract lexical form (pattern). The morphological processor, however, uses pattern matching instead of finite state analysis.

One of the most important components of the morphological processor is the electronic dictionary (Nedobejkine N. 1990, Courtois B. 1990). The building of such a dictionary requires many years. In order to decrease this time robust techniques are implemented. Environments like INTEX that allow users to create and maintain their own lexical data are created (Silberztein M. 1993). Also INTEX allows linguists to describe a natural language from its alphabet, up to the syntactic level, that comes with built-in, large-coverage dictionaries and grammars, and can parse texts of several million words in real time.

A morphological processor for Bulgarian has been built, based on the common properties of inflectional morphology (Totkov G. et al. 1988). It has been implemented in some multilingual environments for natural language generation (AGILE "Automatic Generation of Instructions in Languages of Eastern Europe" Inco-Copernicus PL961104) as well as for anaphora resolution (Collaborative project with the University of Wolverhampton, UK). Linguistic tools for researches on the Bulgarian language have been created (Krushkov Hr. 1999, Tanev Ch. & Krushkov Hr. 2000). Unfortunately the previous version of the processor didn't process proper nouns. The lack of proper nouns in the dictionary was a great disadvantage, taking into account the frequent use of proper nouns in the text corpora.

Bulgarian belongs to the group of the inflectional languages. Bulgarian inflection is described as a number of grammatical rules. A classification of Bulgarian inflection in view of the mentioned rules and the grammatical features of the words has been made (Krustev B. 1984). There are 187 different inflectional classes in that classification divided into parts of speech. In this classification the proper nouns are not presented. The treatment of proper nouns is not widely presented in other languages too. This paper is an attempt to extend the "morphological territory" through systematisation and taxonomy of lexical units which have not been thoroughly investigated till now as proper nouns are. The problem is that the morphological phenomena of proper nouns are more complex than of common nouns. How can we define the paradigm of a proper noun, is it an inflectional or derivational product etc. Possessive forms function as adjectives in Bulgarian sentences. That's why the paradigm is a derivational product. Exploration of derivational rules is needed. The main question is whether they are as regular as inflectional rules in order to use the morphological processor, which manipulates common nouns. Problems of the same type, connected with the particular status of proper nouns appear in other inflectional languages (Maurel D. et al. 1995).

In this paper we describe a classification of Bulgarian proper nouns. This kind of classification is made for the first time in the Bulgarian language. A set of 26 classes – new for Bulgarian grammar is presented, which enforce the power of the "traditional" morphological engine through manipulating the

whole system of proper nouns in the Bulgarian language. An electronic dictionary of proper nouns has been created. In the dictionary only one lexical form (pattern) has been stored, taken as the representative for all the various forms of a paradigm. The generation rules are not described using a description language but they are automatically extracted from the paradigm. Later the dictionary of proper nouns is included in the electronic morphological dictionary.

1. CLASSES OF PROPER NOUNS

The inflection of the Bulgarian proper nouns is described as a number of grammatical rules. Some of these rules are listed in Table 1:

Rule	Example – Cyrillic	Example – Latin
<i>a/e</i>	<i>Милка – Милке</i>	<i>Milka – Milke</i>
reduction of <i>e</i>	<i>Павел – Павлов</i>	<i>Pavel – Pavlov</i>
adding of a suffix	<i>Иван – Иванов</i>	<i>Ivan – Ivanov</i>

Table 1. Some inflectional rules

In this example '*Milka*' is a woman's name, '*Milke*' is the same proper noun in vocative case, '*Pavel*' and '*Ivan*' are men's names, '*Pavlov*' and '*Ivanov*' are either possessive forms or family names deriving from them. We present a classification of the Bulgarian proper nouns in view of the mentioned rules. There are 22 different inflectional and 4 non-inflectional classes in that classification divided as follows: 15 for these, deriving from Bulgarian men's names, 5 for proper nouns deriving from women's names, 2 for family names, 2 for adopted foreign names, 1 for geographic names and 1 for other names. Every Bulgarian proper noun can be classified as a member of some of these classes.

From a mathematical point of view Bulgarian proper nouns are divided into disjoint classes of equivalence. Every class has a unique machine number for identification and a list of rules for the generation of the paradigm.

Two proper nouns are in the same class if their paradigms are generated in the same way. The paradigm is described as a list of wordforms with definite grammatical features for each of them. Every wordform also has a number. Two wordforms with equal numbers have the same grammatical features.

For example (Table 2) in the paradigm of proper nouns, wordform num. 1 is the base form, wordform num. 2 is the vocative form, wordform num. 3 is a possessive form, masc. etc. (see also table 4). This form also may refer to a man's family name. For all proper nouns wordform num. 1 is the base (citation) form.

wordform number	grammatical features of the forms
1	base form
2	vocative form
3	possessive form, masc. / man's family name
4	plural
5	plural, definite form
6	possessive fem. / woman's family name
7	possessive fem., definite form
8	possessive neuter
9	possessive neuter, definite form
10	possessive pl. / family name in plural
11	possessive pl., definite form
12	possessive masc., short definite form
13	possessive masc., full definite form

Table 2. Wordform features in the paradigm of proper nouns.

The description of the proposed 22 morphological classes of proper nouns follows.

2. DESCRIPTION OF CLASSES

All the inflectional classes are described, as well as the formal criteria for automatic classification of proper nouns to the particular class.

2.1. Names, derived from man's name

Class number 1:

Men's names ending in a consonant letter belong to this class.

Тодор, Любен, Стоимен, Явор

Todor, Lyuben, Stoimen, Yavor

The suffixes of derived forms (numbered from 2 to 13) are:

-е, -ов, -овци, -овците, -ова, -овата, -ово, -овото, -ови, -овите, -овия, -овият.

-е, -ов, -овтци, -овтците, -ова, -овата, -ово, -овото, -ови, -овите, -овия, -овият.

For example the forms derived from the name *Тодор (Todor)* are:

Тодоре, Тодоров, Тодоровци, Тодоровците, Тодорова, Тодоровата, Тодорово, Тодоровото, Тодорови, Тодоровите, Тодоровия, Тодоровият.

Todore, Todorov, Todorovtci, Todorovtcite, Todorova, Todorovata, Todorovo, Todorovoto, Todorovi, Todorovite, Todoroviya, Todoroviyat

Class number 2:

Most of the men's names ending in *-o* (*Христо – Hristo*), but not in *-чо* (*tcho*) or *-шо* (*sho*) belong to this class. The vocative form and the base form are the same.

Class number 3:

This class comprises these men's names with ending *-u* (*-i*), which don't lose the ending after the derivation. In this class the base form also coincides with the vocative form. The wordforms are suffixed with:

-ев, -евци, -евците, -ева, -евата, -ево, -евото, -еви, -евите, -евия, -евият.

-ev, -evtci, -evtcite, -eva, -evata, -evo, -evoto, -evi, -evite, -eviya, -eviyat.

A representative of this class is the name *Георги* (*Georgi*):

Георги: Георгиев, Георгиевци, Георгиевците, Георгиева, Георгиевата, Георгиево, Георгиевото, Георгиеви, Георгиевите, Георгиевия, Георгиевият.

Georgi: Georgiev, Georgievtcı, Georgievtcite, Georgieva, Georgievata, Georgievo, Georgievoto, Georgievi, Georgievite, Georgieviya, Georgieviyat.

Class number 4:

The names of this class end in *-чо* (*tcho*) or *-шо* (*sho*) e.g. *Стойчо* (*Stojtcho*). The last letter *-o* is dropped out when the other forms are derived. The vocative form is the same as the base one.

Class number 5:

This class is obtained from class number 3. The difference is that the ending *-u* (*-i*) of the base form is reduced. A representative of this class is the name *Добри* (*Dobri*).

Class number 6:

A few men's names with ending *-a* belong to this class (*Никола – Nikola*). This ending is typical of women's names. It is omitted when other forms are derived.

Class number 7:

Here the names have an ending *-ьо* (*yo* after a consonant) or *-йо* (*yo* after a vowel) e.g. *Кольо* (*Kolyo*). The ending is omitted, except in forms 4 and 5.

Class number 8:

Names with the ending *-ър* (*-yr*) belong to this class. In other forms only the letter 'ъ' (*y*) is reduced. The wordforms suffixes are the same as in the class number 1. The paradigm of the name *Петър* (*Petyr*) is presented in table 4.

Class number 9:

This class is close to class number 1. Only the possessive forms have other inflectional morphemes, in which the morpheme *-ев* (*-ev*) is used instead of *-ов* (*-ov*).

Class number 10:

These are names with last letter *-я* (*ya*), which is reduced at the derivational process. The wordforms suffixes from class number 3 are the same.

Class number 11:

The only difference between this class and class number 1 is in the vocative form. In this class this form coincide with the base form. It comprises a lot of old Bulgarian names like *Аспарух* (*Asparuh*), *Омуртаг* (*Omurtag*) etc.

Class number 12:

This class is close to class 3. The difference is in the ending of the base form *-ъ* (short *i*) e.g. *Матеъ* – (*Matei*) unlike the ending *-у* (*-i*) of class 3. Another difference is that here the ending *-ъ* (short *i*) is reduced.

Class number 13:

This class is close to class 2. The difference is in the ending of the base form *-е* unlike the ending *-о* of class 2.

Class number 14:

The base form in this class ends in *-у* (*-i*) like in the class 5, which also is reduced. But the suffixes added to the derived forms are these from class number 6.

Class number 15:

Until now only one representative of this class is found. This is the name *Павел* (*Pavel*). Reduction of 'е' at the derivational process appear, except in the vocative form:

Павел: Павеле, Павлов, Павловци, Павловците, Павлова, Павловата, Павлово, Павловото, Павлови, Павловите, Павловия, Павловият.

Pavel: Pavele, Pavlov, Pavlovtsi, Pavlovtcite, Pavlova, Pavlovata, Pavlovo, Pavlovoto, Pavlovi, Pavlovite, Pavloviya, Pavloviyat.

2.2. Names, derived from woman's name

The next 5 classes are for women's names.

Class number 16:

All names with reducible ending *-а* (but not *-ка*) belong to this class. The inflectional morphemes are:

-о, -и, -ин, -ите, -ина, -ината, -ино, -иното, -ини, -ините, -иния, -иният.

-о, -і, -in, -ite, -ina, -inata, -ino, -inoto, -ini, -inite, -iniya, -iniyat.

The paradigm of *Елена* (*Elena*) is:

Елена: Елено, Еленин, Елени, Елените, Еленина, Еленината, Еленино, Елениното, Еленини, Еленините, Елениния, Елениният.

Elena: Eleno, Elenin, Eleni, Elenite, Elenina, Eleninata, Elenini, Eleninoto, Elenini, Eleninite, Eleniniya, Eleniniyat.

Class number 17:

Like the previous. Only the vocative form obtains the ending -e.

Class number 18:

All names with reducible ending -я (-ya) e.g. *Галя (Galya)* belong to this class. The inflectional morphemes are these from the other two classes. Only the vocative form and the base form are the same.

Class number 19:

The women's names with ending -и (-i). Their plural forms obtain ending -ма (ta).

Class number 20:

All other women's names are in this class. They end in a consonant.

2.3. Names, derived from family name

Family names without corresponding base forms exist. The next two classes comprise such family names. Their paradigms miss the first two forms.

Class number 21:

Here are family names with suffix -ов (-ov) or -ев (-ev). The first two forms are missing. The inflectional morphemes are:

-ци, -ците, -а, -ата, -о, -ото, -и, -ите, -ия, -ият.

-тци, -тците, -а, -ата, -о, -ото, -и, -ите, -ия, -ият.

The paradigm of *Коларов (Kolarov)* is:

Коларов: Коларовци, Коларовците, Коларова, Коларовата, Коларово, Коларовото, Коларови, Коларовите, Коларовия, Коларовият.

Kolarov: Kolarovtci, Kolarovtcite, Kolarova, Kolarovata, Kolarovo, Kolarovoto, Kolarovi, Kolarovite, Kolaroviya, Kolaroviyat.

Class number 22:

The family names that end in -ски (-ski) belong to this class. The last letter -и (-i) is reduced. The paradigm misses the first 5 forms. The inflectional morphemes are:

-а, -ата, -о, -ото, -и, -ите, -ия, -ият.

-а, -ата, -о, -ото, -и, -ите, -ия, -ият.

The paradigm of *Матански (Matanski)* is:

Матански: Матанска, Матанската, Матанско, Матанското, Матански, Матанските, Матанския, Матанският.

Matanski: Matanska, Matanskata, Matansko, Matanskoto, Matanski, Matanskite, Matanskiya, Matanskiyat.

The other classes are not inflectional. Every member of these classes has only one form. These classes comprise adopted foreign names and geographical names.

An approach for automatic morphological generation and analysis is investigated, based on the presented classification.

3. MORPHOLOGICAL GENERATION

For every word a pattern is built up. The pattern and the inflectional class number determine the paradigm of this word. The pattern defines which letters are constant in all wordforms in the paradigm of the word and which are variable. The variable letters are marked with '*' in the pattern.

3.1. Automatic acquisition of the pattern from the paradigm

The pattern can be automatically extracted from the paradigm following this simple algorithm:

1. Pattern:=''.
2. Extract the initial letter of the base-form.
3. **If** the letter is constant at the corresponding place for all wordforms in the paradigm
 then pattern:=pattern+letter
 else pattern:=pattern+ '*'.
4. **If** there are no more letters in the base form
 then end
 else extract the next letter from the base form and go to 3.

An example of this algorithm for the pattern extraction (Table 3) from the paradigm of the name 'Петър' (Peter – Table 4) follows:

Step	Action	Pattern formation
1	Pattern:='';	''
2	letter:='P'	
3	'P' is constant → Pattern=''+ 'P'	'P'
4	Extract the next letter → letter='e'	
3	'e' is constant → Pattern='P'+ 'e'	'Pe'
4	Extract the next letter → letter='t'	
3	't' is constant → Pattern='Pe'+ 't'	'Pet'
4	Extract the next letter → letter='y'	
3	'y' is not constant → Pattern='Pet'+ '*'	'Pet*'
4	Extract the next letter → letter='r'	
3	'r' is constant → Pattern='Pet*'+ 'r'	'Pet*r'
4	There are no more letters in the base form → end.	

Table 3. An example of the pattern extraction

number/wordform	grammatical features	translation
1. <i>Петър</i> (<i>Petyr</i>)	base form (man's name)	Peter
2. <i>Петре</i> (<i>Petre</i>)	vocative form	Peter
3. <i>Петров</i> (<i>Petrov</i>)	possessive form, masc. / man's family name	Peter's / Peter
4. <i>Петровци</i> (<i>Petrovtci</i>)	plural	Peters
5. <i>Петровците</i> (<i>Petrovtcite</i>)	plural definite form	The Peters
6. <i>Петрова</i> (<i>Petrova</i>)	possessive fem. form / woman's family name	Peter's / Peter
7. <i>Петровата</i> (<i>Petrovata</i>)	possessive fem. definite form	The Peter's
8. <i>Петрово</i> (<i>Petrovo</i>)	possessive neuter form	Peter's
9. <i>Петровото</i> (<i>Petrovoto</i>)	possessive neuter definite form	The Peter's
10. <i>Петрови</i> (<i>Petrovi</i>)	possessive pl. form / family name in plural	Peters' / Peters
11. <i>Петровите</i> (<i>Petrovite</i>)	possessive pl. definite form	The Peters'
12. <i>Петровия</i> (<i>Petroviya</i>)	possessive masc. short definite form	The Peter's
13. <i>Петровият</i> (<i>Petroviyat</i>)	possessive masc. full definite form	The Peter's

Table 4. An example of a paradigm

The pattern of the word '*Петър*' (*Petyr*) is '*Пет*р*' (*Pet*r*). All other words from the same inflectional class '*Димитър*' (*Dimityr*), '*Александър*' (*Aleksandyr*), with the following patterns: '*Димит*р*' (*Dimit*r*), '*Александ*р*' (*Aleksand*r*) have a variable letter – the last vowel in the pattern.

The pattern involves some features important for morphological analysis as follows:

1. The length (number of letters except asterisks) of the pattern is less than or equal to the length of every wordform of the paradigm generated from that pattern.

2. The pattern matches the beginning of every wordform with a full coincidence of the constant letters, taking into account their order in the base form.

3.2. Automatic extraction of synthesis rules

The rules for the wordform generation are of two types:

1. Replacing the "*" with a letter (including the zero character – "").
2. Inflectional rules (adding suffixes).

For example the rules for the above-mentioned inflectional class are the following:

number/wordform	rules – Cyrillic	rules – Latin
1.Петър (<i>Petyr</i>)	*/'Ъ'	*/'y'
2.Петре (<i>Petre</i>)	*/' + 'e'	*/' + 'e'
3.Петров (<i>Petrov</i>)	*/' + 'os'	*/' + 'ov'
4.Петровци (<i>Petrovtci</i>)	*/' + 'os' + 'cu'	*/' + 'ov' + 'tci'
5.Петровците (<i>Petrovtcite</i>)	*/' + 'os' + 'cu' + 'me'	*/' + 'ov' + 'tci' + 'te'
6.Петрова (<i>Petrova</i>)	*/' + 'os' + 'a'	*/' + 'ov' + 'a'
7.Петровата (<i>Petrovata</i>)	*/' + 'os' + 'a' + 'ma'	*/' + 'ov' + 'a' + 'ta'
8.Петрово (<i>Petrovo</i>)	*/' + 'os' + 'o'	*/' + 'ov' + 'o'
9.Петровото (<i>Petrovoto</i>)	*/' + 'os' + 'o' + 'mo'	*/' + 'ov' + 'o' + 'to'
10.Петрови (<i>Petrovi</i>)	*/' + 'os' + 'u'	*/' + 'ov' + 'i'
11.Петровите (<i>Petrovite</i>)	*/' + 'os' + 'u' + 'me'	*/' + 'ov' + 'i' + 'te'
12.Петровия (<i>Petroviya</i>)	*/' + 'os' + 'u' + 'ya'	*/' + 'ov' + 'i' + 'ya'
13.Петровият (<i>Petroviyat</i>)	*/' + 'os' + 'u' + 'yat'	*/' + 'ov' + 'i' + 'yat'

Table 5. The generation rules for class number 8

The pattern and the inflectional class number incorporate information for the whole paradigm of a particular word. The inflectional class involves for every wordform a list of letters (including the empty one – ‘’) for replacing the symbol ‘*’ of the pattern and inflectional morphemes for adding the pattern.

The morphological generation of the paradigm is based on the following simple mechanism: every wordform can be constructed from the pattern operating with the rules of replacing (* / letter) and adding (+ morpheme), described after the wordform number. Once extracted, the rules for a member of some inflectional class are the same for all other members of this class.

4. MORPHOLOGICAL ANALYSIS

The goal of the automatic morphological analysis is to perform automatically a morphological classification of an arbitrary wordform. It includes identifying the base form of the word, its grammatical features and to which inflectional class it belongs. In case of homonyms (when the wordform belongs to more than one inflectional classes and has different grammatical features) all possible classes must be found.

The electronic dictionary consists of pairs <word-pattern, inflectional class number>. When an arbitrary wordform has to be classified the analyser looks up a matching word-pattern in the dictionary. If such a pattern has been found, using the second part of the entry pair (inflectional class number) the rules are extracted from the generation table. A paradigm is generated from this pattern on the basis of these rules. If the analysed word coincides with a wordform from the generated paradigm it obtains the grammatical features of that wordform. In such a way the word is morphologically completely determined. It obtains 2 formal features:

1. An inflectional class number
2. A wordform number in the paradigm of that class.

Moreover – the first wordform of the generated paradigm is the base-form of the analysed word. The inflectional class number determines the subclass of the analysed word. For example if that number is from 1 to 15 that means the proper noun derived from man's name, etc.

5. AUTOMATIC CREATION OF A DICTIONARY OF PROPER NOUNS

Three sources are used for the creation of a dictionary of proper nouns. The first is Spelling Dictionary of Contemporary Bulgarian Language (Georgieva El. 1983). All words with capitalised first letter are extracted. The second one is Frequency Dictionary of Bulgarian Proper Nouns (Stankov V. 1984). Proper nouns with frequency 300 and more from this dictionary are added. All words from both sources are in base form. Newspaper corpora are the last source. The words with capitalised first letter are extracted except the first word in every sentence and names in double quotes (in Bulgarian – names of firms, journals, newspapers etc.), which always begin with such a letter. In this case the proper nouns are divided into two types – base forms and derived forms. The automatic recognition of derived forms is made by exploring the suffixes presented in the classification. Later on, separated derived forms serve to approve the hypotheses made by the algorithm for automatic classification.

The number of wordforms in the all corpora is 369888. We extracted 23284 wordforms with capitalised first letter, 6757 of which are different. The morphological processor recognised 2492 like common nouns, adjectives etc. In this moment the processor worked without proper nouns in its electronic dictionary. From the remaining 4265 forms we dropped out those with less than two citations (2441). Afterwards, 1780 wordforms with length (number of letters) 3 and more were explored. As a result, 592 from them were non-inflected names: 94 with foreign origin (82 men's names, 12 women's names) and 498 geographical names (cities, countries, mountains etc.); 105 – other names and misspelt words. The base forms (756) were separated from the remaining 1083 forms.

The next step is to determine a class number of every base form using the algorithm for automatic classification. After running the program based on the algorithm, a hypothetical class number is attached to every base form. A lexicographer checked the hypotheses. Some classes are additionally divided into new classes. For example men's names with ending '-u' (-i) may generate possessive forms in different ways:

Георџи – Георџиев
Georgi – Georgiev
Добри – Добрев (reduction of '-u')
Dobri – Dobrev (reduction of '-i')

Finally 404 base forms (from among 621 inflected forms) for men's names, 214 (from among 253) – for women's names as well as 138 (from among 209) – for family names are classified after extraction from the corpora.

Table 6 demonstrates the distribution of the proper nouns in the electronic dictionary by classes. The last four classes are non-inflectional, comprising as follows: 23 – adopted foreign women’s names; 24 – adopted foreign men’s names; 25 – geographical names; 26 – other names.

Class Number	Number Of Proper Nouns
1	299
2	162
3	23
4	98
5	17
6	7
7	69
8	3
9	2
10	3
11	4
12	20
13	1
14	1
15	1
16	477
17	284
18	78
19	11
20	2
21	145
22	25
23	12
24	82
25	1100
26	27
Total in all	2953

Table 6. Distribution of proper nouns by classes in the electronic dictionary.

CONCLUSION

We have presented the process of building a Bulgarian proper noun dictionary. This process begins with the classification of Bulgarian proper nouns. The classification is close to the same of common nouns. Though in the proper noun paradigm inflection and derivation are mixed, derivational rules are as much productive and regular as inflectional rules are. In such a way the treatment of all nouns lays on the common base. Finally the

electronic dictionary of proper nouns comprises 2953 base forms which produce over 24000 inflected forms. It is added to the general morphological dictionary with class numbers from 201 to 226, which compressed size is 240 KB. Before analysis the dictionary is loaded into the memory. This allows performing the analysis with a maximum speed. The dictionary is distributed by ELRA (European Language Resources Agency). Remote access to the dictionary is possible on the WEB address:

<http://www.pu.acad.bg/dcs/morphe.htm>

The used software is Borland Pascal 7.0/Delphi working on MS DOS and WINDOWS 95 platforms. A dynamic link library has been created in order to be available for different WINDOWS applications.

Filling the gap in the Bulgarian electronic dictionary we contribute in the mentioned (in the introduction) international projects for natural language generation as well as for anaphora resolution. Linguistic software for WINWORD has been developed for natural language researches. It also uses the specified DLL.

Obviously a classification of this kind can be applied to geographical names. This is an issue of further research. Of course, the methodology we have developed for investigating the morphology of Bulgarian proper nouns may also be applied to any other Slavic language as well as to inflectional languages.

ACKNOWLEDGMENTS

I am thankful to Prof. Georgi Totkov from the Department of Computer Science at the University of Plovdiv, to Dr. Iliyana Krapova from the Department of Bulgarian at the same University as well as to the anonymous reviewers for many useful notes, comments and suggestions.

REFERENCES

- COURTOIS, Blandine (1990): "Dictionnaires électroniques du français", *Langues française*, n87, 11-22, Paris, Larousse.
- GEORGIEVA, Elena (1983): *Spelling Dictionary of Contemporary Bulgarian Language* (in Bulgarian), Sofia, Bulgarian Academy of Science.
- KOSKENNIEMI, Kimmo (1983): *Two-level morphology: a general computational model for word-form recognition and production*. Publication n11. Helsinki, University of Helsinki, Department of General Linguistics.
- KRUSHKOV Hristo (1999): "Development of linguistic software for WORD 7.0", in *Proc. 28-th spring conference of the Union of Bulgarian Mathematicians*, Montana, pp.199-203.
- KRUSTEV Borimir (1984): *Bulgarian Morphology in 187 tables* (in Bulgarian), Sofia, Nauka I Izkustvo.
- MAUREL Denis, LEDUC B., COURTOIS Blandine (1995): "Vers la constitution d'un dictionnaire électronique des noms propres", *Linguisticae Investigationes*, volume 19:2, p. 355-368.

- NEDOBEJKINE Nicolas (1990): "Représentation des informations lexicales dans les dictionnaires électroniques", *TA information*, v31, n1, 5-15.
- SILBERZTEIN Max (1993): *Dictionnaires électroniques et analyse automatique de textes - Le système INTEX*, Paris, Masson.
- SPROAT Richard (1992) *Morphology and Computation*, Cambridge, Massachusetts, The MIT Press.
- STANKOV Vasil (1984): *Frequency Dictionary of Bulgarian Proper Nouns* (in Bulgarian), Sofia, Bulgarian Academy of Science.
- TANEV Christo, KRUSHKOV Hristo (2000): Hr., "Language processing tools for Bulgarian", in *Proc. ACIDCA'2000 - Corpora and Natural Language Processing*, Monastir, Tunisia, pp.221-227.
- TOTKOV Georgi, KRUSHKOV Hristo, KRUSHKOVA Mariana (1988): *Formalisation of Bulgarian Language and the Development of a Linguistic Processor (Morphology)* (in Bulgarian), *Travaux scientifiques d'Université de Plovdiv*, vol.26, fasc.3-Mathématique.
- TOTKOV Georgi (1990): "Robust Methods for Automated Analysis of Bulgarian Texts and the Development of a Linguistic Processor" (in Bulgarian), in *Proc. 19th spring conference of the Union of Bulgarian Mathematicians*, pp. 295-303.
- TOTKOV Georgi (1991): "The Development of a Linguistic Processor: problems, results, future" (in Bulgarian), in *Proc. 20th spring conference of the Union of Bulgarian Mathematicians*, pp. 43-50.